Prémio Arquivo.pt

Identificação

- Título: KairosNews
- Área temática: Ciência de dados, Tecnologias da informação
- Candidatos: Christopher Abreu, Francisco Alves, Quintino Fernandes
- Email: christopherabreubusiness@gmail.com ;aa49@live.com.pt; quintinoferprof@gmail.com

Descrição do Trabalho

KairosNews é um projeto que nasceu a partir de uma tese de mestrado, sendo o framework e os dados desta adaptados para uma plataforma web e aplicação que permite o acesso a uma cápsula do tempo na forma de um resumo de notícias passadas, de acordo com especificações do utilizador.

Isto permite ao utilizador pesquisar notícias relacionadas com qualquer frase inserida, podendo restringir por intervalo de tempo e por tópico, recebendo no fim um pequeno resumo combinado das 5 notícias mais relevantes e os links das mesmas. O projeto desenvolvido é composto por duas componentes:

1ª Componente: Pesquisa de notícias:

Uso de uma dataset desenvolvido por nós em *PostgreSQL*, com mais de 200 mil artigos de 2020 a 2024 que foram retirados do Arquivo.pt usando a sua API, hosted numa database online e disponibilizado livremente.
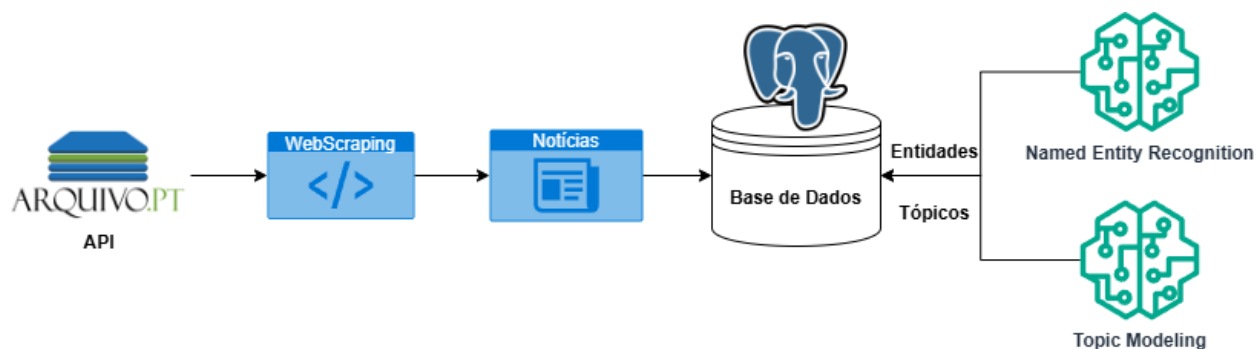
Figura 1 Criação do Dataset

Este dataset é composto por:

	Descrição
url	Link para a notícia arquivada no Arquivo.pt
Título	Título da notícia
Contéudo	Corpo da notícia
Data	Data da notícia
Tópico	Tópico da notícia, classificado de acordo com um modelo de Tópic Modeling
Embedding	Vector gerado por um modelo Transformer de linguagem
Entidades	Entidades (ex: Pessoa, Localidade, Organização, Misc.) detetadas por um modelo de NER

Para descobrir os 5 artigos mais relevantes é feita uma pesquisa em 3 fases: primeiramente, é feita uma filtragem inicial de acordo com o intervalo de tempo e tópico que for escolhido pelo utilizador; em segundo lugar, a frase inserida pelo utilizador é analisada por um modelo de *Named Entity Recognition(NER)* e as entidades são extraídas e comparadas com os artigos encontrados no primeiro passo, selecionando os artigos onde estas aparecem; por último, a frase inserida pelo utilizador é transformada num vetor de 384 dimensões usando um modelo *Transformer* e é feita uma pesquisa semântica nos artigos filtrados anteriormente. Esta pesquisa semântica consiste em calcular a distância entre o vetor da frase inserida pelo utilizador e os vetores dos artigos filtrados, selecionando os 5 artigos com a distância menor.
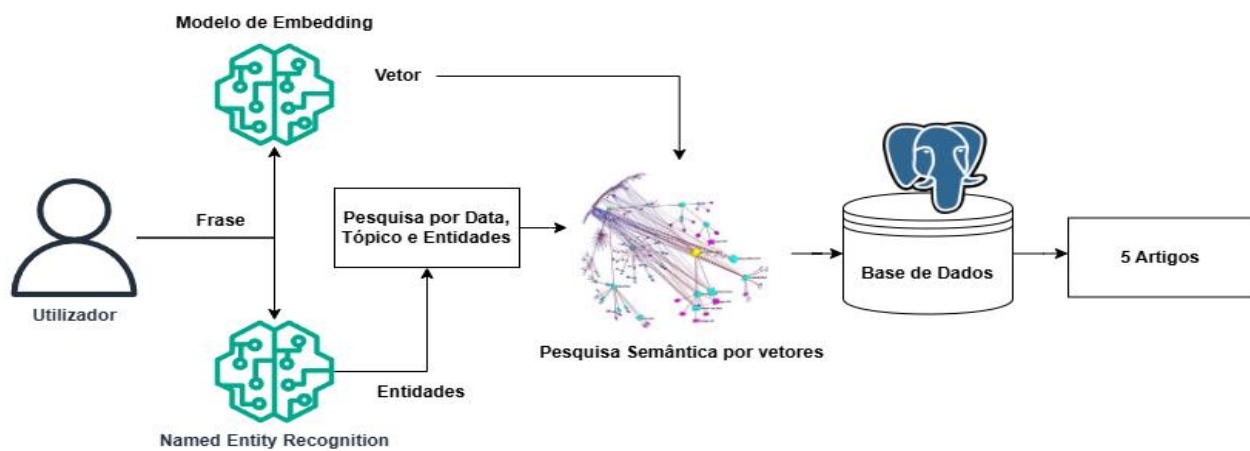
Figura 2 Pesquisa das Notícias

2ª Componente: Sumarização:

Para o sumário dos 5 artigos de notícias foi desenvolvida uma framework, utilizando dois tipos diferentes de sumarização. A primeira parte separa os 5 artigos em frases, sendo estas depois transformadas em vetores novamente por um modelo *Transformer*. Posteriomente, é criado um gráfico de arestas e vértices, relacionando as distâncias dos vetores das frases entre si. As 10 frases mais próximas/relacionadas são extraídas, compondo um resumo inicial.

A segunda parte, a sumarização abstrativa, é composta por um *Large Languague Model* de 223 milhões de parâmetros (PTT5), que aceita as 10 frases extraídas anteriormente e as transforma num pequeno parágrafo conciso.
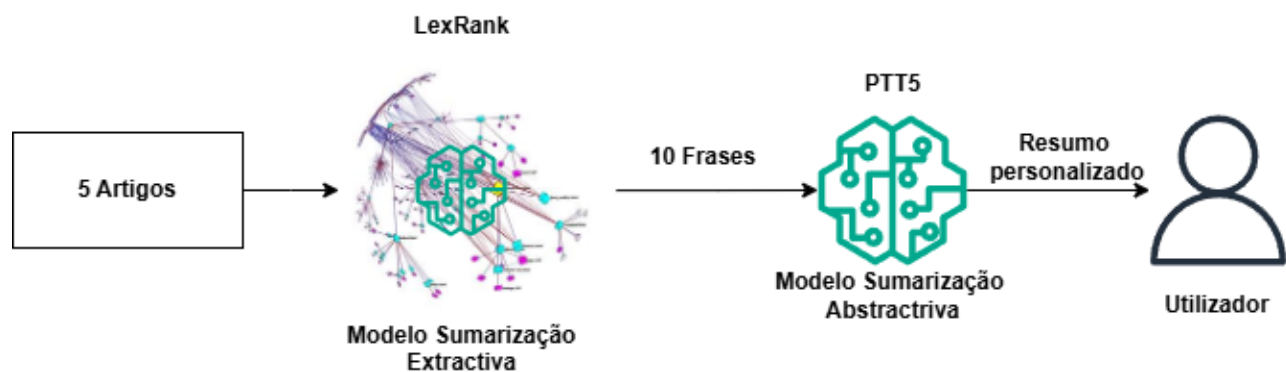
Figura 3 Framework de Sumarização

Objetivos

Este trabalho visa contribuir para a resolução de três problemas que afetam o cidadão moderno:

Em primeiro lugar, constata-se nos dias correntes uma sobrecarga da informação disponível nos vários meios de comunicação. Numa realidade onde existe uma competição pela atenção dos leitores, verifica-se que a tarefa destes se manterem atualizados é progressivamente um processo mais complexo e temporalmente exigente.

Em segundo lugar, com o advento da inteligência artificial, há um empoderamento de indivíduos e minorias para alavancarem estas tecnologias para fins pessoais, produzindo "notícias" que, por um lado, não respeitam os valores que regem o bom jornalismo (nomeadamente, a verdade e

precisão) e, por outro, visam frequentemente uma polarização dos leitores. Esta polarização serve comummente os interesses particulares dos seus criadores, não almejando o bem comum do ponto de vista social de informar devidamente a população geral.

Por último, face à velocidade crescente a que a realidade se vive, comparativamente a tempos mais antigos, emerge uma necessidade por muitos de encontrar informação que, muito embora seja de qualidade e fiável, possa igualmente ser consumida em tempo útil. Visamos resolver esta necessidade, mas sem negar ao leitor a possibilidade de aprofundar o tópico em fontes reputáveis, se assim desejado pelo mesmo.

Resultados Atingidos

Como foi referido anteriormente, foi criado e disponibilizado o dataset completamente processado e pronto para recriar o nosso trabalho ou para outras utilidades no mesmo âmbito. Foi também atingido o resultado de hospedar online tanto o dataset como framework, para poderem ser usados de uma forma eficiente com a plataforma web criada por nós, de fácil e intuitiva utilização, também disponibilizada ao público geral.

Como resultado principal, atingimos o objetivo de desenvolver um framework de sumarização comparável aos modelos de texto de última geração como o *ChatGPT* e *Deepseek*, com um framework de tamanho 16 mil vezes menor, poupando potência computacional e fazendo que o nosso projeto seja facilmente replicável e adaptável sem a necessidade de altos requisitos de hardware. O framework foi também disponibilizado para uma fácil recriação do nosso projeto

Figura 4 O nosso website

Originalidade e caráter inovador

Tanto quanto sabemos, não existe nos dias correntes uma ferramenta que permita a síntese das notícias portuguesas de acordo com os critérios gerados pelo leitor e adaptados a um enquadramento temporal escolhido pelo mesmo. Acresce a particularidade do uso da base de dados do Arquivo.pt, garantindo assim a fiabilidade das notícias resumidas.

Adicionalmente, a inclusão das referências para as notícias mais relevantes (referentes à pesquisa do leitor) torna este site não apenas uma ferramenta de resumo, mas também um motor de pesquisa aperfeiçoado para os que assim o desejarem.

Impacto social (aplicação e utilidade social)

O nosso *website* pretende enriquecer o debate social, permitindo aos leitores uma atualização para os tópicos que desejem, em tempo útil e com base em informação de confiança. Desde o cidadão que, ao acordar, deseja um resumo de um tema que lhe seja importante, passando pelos debates entre amigos dos quais se levantam as questões específicas que exigiriam uma procura morosa, até ao leitor que procura a compreensão de um tema dentro de um período do passado, pretende-se que o KairosNews seja uma ferramenta de auxílio que eleve a qualidade do acesso à informação para o cidadão comum na sua vida habitual.

Ao disponibilizar a aplicação para os sistemas de *Android* e *IOS*, além do *website* criado, esperamos que a versatilidade dos usos possíveis seja acompanhada por facilidade e universalidade no seu acesso. Esperamos que KairosNews seja acessível por todos, em qualquer lugar e em qualquer sítio, bastando apenas uma ligação à *internet*.

Impacto científico (aplicação e utilidade científica)

Do ponto de vista científico, e realçando a possibilidade de utilização do website como motor de pesquisa, esperamos que o KairosNews traga valor aos investigadores no processo de procura de informação. Através da nossa *webapp* é possível realizar perguntas tão complexas e específicas quanto desejado, sendo fornecido o *link* para as notícias mais relevantes referentes a essa pesquisa. Acreditamos que o KairosNews se revela um instrumento que traz simplicidade a um processo de pesquisa de elevada qualidade.

Adicionalmente, o nosso código e dataset são disponibilizados livremente. Com este livre acesso esperamos que este seja reproduzido por quem estiver interessado, permitindo também que seja

uma base a partir da qual possam ser realizadas alterações mais específicas, sem que seja necessário todo o trabalho que tivémos de realizar.

Relevância da utilização do Arquivo.pt

A construção deste trabalho assentou no acesso ao API do Arquivo.pt. Sem a disponibilização desta, não seria possível qualquer processo ou uso que o KairosNews realiza e fornece. Numa tentativa hipotética de conceber uma alternativa, chega-se à conclusão imediata da sua impossibilidade, visto que a) a maioria dos jornais não guarda/facilita o acesso às suas notícias passadas, b) haveria entraves de ordem legal no acesso e utilização das mesmas e c) seria impossível de realizar um projeto exequível atendendo aos custos financeiros necessários. Consideramos, assim, que o Arquivo.pt é indispensável e fundamental para a realização deste projeto.

As notícias disponibilizadas pelo Arquivo.pt encontram-se integralmente preservadas, sendo disponibilizadas ao leitor através de uma hiperligação.

Comentários adicionais

Kairos vem do grego antigo *καιρός*, que se refere ao tempo oportuno, o momento fugaz em que a ação deve ser tomada. O nosso nome advém da crença que é impossível separar os factos dos momentos em que estes acontecem, e que é o contexto temporal que muitas vezes explica as ações praticadas. Surge daí a nossa contextualização temporal, permitindo a escolha ao leitor do período que quer ver analisado.

Recursos complementares

- O website do arquivo: https://arquivo.pt/
- A nossa plataforma web: https://kairos-news.expo.app/
- Os nossos repositórios de código aberto:

  https://github.com/0edon/KairosNews - Framework de sumarização

  https://github.com/cfa911/KairosNews - Website